Executive Summary

Despite some apparent differences, biology and information technology (IT) have much in common. They are two of the most rapidly changing fields today—the former because of enormous influxes of new, highly heterogeneous data, and the latter because of exponentially decreasing price-performance ratios. They both deal with entities of astounding complexity (organisms in the case of biology, networks and computer systems in the case of information technology), although in the IT context, the significance of the constituent connections and components is much better understood than in the biological context. Also, they both have profound and revolutionary implications for science and society. Biological science and technology have the potential to contribute strongly to society in improving human health and well-being. The potential impacts include earlier diagnoses and more powerful treatments for diseases, rapid environmental cleanup, and more robust food production. Computing and information technology enable human beings to acquire, store, process, and interpret enormous amounts of information that continue to underpin much of modern society.

Against that backdrop, this report considers potential interactions between biology and computing—the "BioComp" interface. To understand better the potential synergies at the BioComp interface and to facilitate the development of new collaborations between the scientific communities in both fields that can better exploit these synergies, the National Research Council established the Committee on Frontiers at the Interface of Computing and Biology. For simplicity, this report uses "computing" to refer to the broad domain encompassed collectively by terms such as computing, computation, modeling and simulation, computer science, computer engineering, informatics, information technology, scientific computing, and computational science. (Analytical techniques without a strong machine-assisted computational dimension are generally excluded from this study, although they are mentioned from time to time when there is an interesting relationship to computing.) Similarly, the report uses the term "21st century biology" to refer to all fields of endeavor in the biological, biochemical, and biomedical sciences.

Obviously, the union of computing with biology results in an extraordinarily broad area of interest. Thus, this report is not intended to be comprehensive in the sense of seeing how every subfield of biology might connect to every topic in computing. Instead, it seeks to sample the intellectual terrain in enough places so as to give the reader a sense of the kinds of activities under way, and its spirit should be understood as "letting a thousand flowers bloom" rather than "identifying the prettiest flowers in the landscape."

COMPUTING'S IMPACT ON BIOLOGY

Twenty-first century biology will integrate a number of diverse intellectual notions. One integration is that of the reductionist and systems approaches—a focus on components of biological systems combined with a focus on interactions among these components. A second integration is that of many distinct strands of biological research: taxonomic studies of many species, the enormous progress in molecular genetics, steps toward understanding the molecular mechanisms of life, and a consideration of biological entities in relationship to their larger environment. A third integration is that computing will become highly relevant to both hypothesis testing and hypothesis generation in empirical work in biology. Finally, 21st century biology will also encompass what is often called discovery science—the enumeration and identification of the components of a biological system independently of any specific hypothesis about how that system functions (a canonical example being the genomic sequencing of various organisms). Twenty-first century biology will embrace the study of an inclusive set of biological entities, their constituent components, the interactions among components, and the consequences of those interactions, from molecules, genes, cells, and organisms to populations and even ecosystems.

How will computing play in 21st century biology? Life scientists have exploited computing for many years in some form or another. Yet what is different today—and will increasingly be so in the future—is that the knowledge of computing needed to address many interesting biological problems can no longer be learned and exploited simply by "hacking" and reading the manuals. Indeed, the kinds and levels of expertise needed to address the most challenging problems of 21st century biology stretch the current state of knowledge of the field—a point that illuminates the importance of real computing research in a biological context.

This report identifies four distinct but interrelated roles of computing for biology.

1. *Computational tools* are artifacts—usually implemented as software but sometimes hardware—that enable biologists to solve very specific and precisely defined problems. Such biologically oriented tools acquire, store, manage, query, and analyze biological data in a myriad of forms and in enormous volume for its complexity. These tools allow biologists to move from the study of individual phenomena to the study of phenomena in a biological context; to move across vast scales of time, space, and organizational complexity; and to utilize properties such as evolutionary conservation to ascertain functional details.
2. *Computational models* are abstractions of biological phenomena implemented as artifacts that can be used to test insights, to make quantitative predictions, and to help interpret experimental data. These models enable biological scientists to understand many types of biological data in context, even in very large volume, and to make model-based predictions that can then be tested empirically. Such models allow biological scientists to tackle difficult problems that could not readily be posed without visualization, rich databases, and new methods for making quantitative predictions. Biological modeling itself has become possible because data are available in unprecedented richness and because computing itself has matured enough to support the analysis of such complexity.
3. A *computational perspective on or metaphor for biology* applies the intellectual constructs of computer science and information technology as ways of coming to grips with the complexity of biological phenomena that can be regarded as performing information processing in different ways. This perspective is a source of information and computing abstractions that can be used to interpret and understand biological mechanisms and function. Because both computing and biology are concerned with function, information and computing abstractions can provide well-understood constructs that can be used to characterize the biological function of interest. Further,

such abstractions may well provide an alternative and more appropriate language and set of abstractions for representing biological interactions, describing biological phenomena, or conceptualizing some characteristics of biological systems.

4. *Cyberinfrastructure and data acquisition* are enabling support technologies for 21st century biology. Cyberinfrastructure—high-end general-purpose computing centers that provide supercomputing capabilities to the community at large; well-curated data repositories that store and make available to all researchers large volumes and many types of biological data; digital libraries that contain the intellectual legacy of biological researchers and provide mechanisms for sharing, annotating, reviewing, and disseminating knowledge in a collaborative context; and high-speed networks that connect geographically distributed computing resources—will become an enabling mechanism for large-scale, data-intensive biological research that is distributed over multiple laboratories and investigators around the world. New data acquisition technologies such as genome sequencers will enable researchers to obtain larger amounts of data of different types and at different scales, and advances in information technology and computing will play key roles in the development of these technologies.

Why is computing in all of these roles needed for 21st century biology? The answer, in a word, is data. The data relevant to 21st century biology are highly heterogeneous in content and format, multimodal in method of collection, multidimensional in time and space, multidisciplinary in creation and analysis, multiscale in organization, international in relevance, and the product of collaborations and sharing. Consider, for example, that biological data may consist of sequences, graphs, geometric information, scalar and vector fields, patterns of organization, constraints, images, scientific prose, and even biological hypotheses and evidence. These data may well be of very high dimension, since data points that might be associated with the behavior of an individual unit must be collected for thousands or tens of thousands of comparable units.

These data are windows into structures of immense complexity. Biological entities (and systems consisting of multiple entities) are sufficiently complex that it may well be impossible for any human being to keep all of the essential elements in his or her head at once; if so, it is likely that computers will be the vessel in which biological theories are held, formed, and evaluated. Furthermore, because of evolution and a long history of environmental accidents that have driven processes of natural selection, biological systems are more properly regarded as engineered entities than as objects whose existence might be predicted on the basis of the first principles of physics, although the evolutionary context means that an artifact is never "finished" and rather has to be evaluated on a continuous basis. The task of understanding thus becomes one of "reverse engineering"—attempting to understand the construction of a device about whose design little is known but from which much indicative empirical data can be extracted.

Twenty-first century biology will be an information science, and it will use computing and information technology as a language and a medium in which to manage the discrete, nonsymmetric, largely nonreducible, unique nature of biological systems and observations. In some ways, computing and information will have a relationship to the language of 21st century biology that is similar to the relationship of calculus to the language of the physical sciences. Computing itself can provide biologists with an alternative, and possibly more appropriate, language and sets of intellectual abstractions for creating models and data representations of higher-order interactions, describing biological phenomena, and conceptualizing some characteristics of biological systems.

BIOLOGY'S IMPACT ON COMPUTING

From the computing side (i.e., for the computer scientist), there is an as-yet-unfulfilled promise that biology may have significant potential to influence computer design, component fabrication, and software. The essential premise is that biological systems possess many qualities that would be desirable in the information technology that humans use. For example, computer and information scientists are looking for ways to make computers more adaptive, reliable, "smarter," faster, and resilient. Biological systems excel at finding and learning good—but not necessarily optimal—solutions to ill-posed problems on time scales short enough to be useful to them. They efficiently store "data," integrate "hardware" and "software," self-correct, and have many other properties that computing and information science might capture in order to achieve its future goals. Especially for areas in which computer science lacks a well-developed theory or analysis (e.g., the behavior of complex systems or robustness), biology may have the most to contribute.

The impact of biology and biological sciences on advances in computing is, however, more speculative than the reverse, because such considerations are, with only a few exceptions, relevant to future outcomes and not to what has been or is already being delivered. Humans understand computing artifacts much better than they do biological organisms, largely because humans have been responsible for the design of computing artifacts. Absent a comparable base of understanding of biological organisms, the historical and contemporary contributions from biology to computing have been largely metaphorical and can be characterized more readily as inspiration, rather than advances having a straightforward or linear impact.

This difference may be one of time scale. Because today's computing already contributes directly in an essential way to advancing biological knowledge, a path for the near-term future can be readily described. Contemporary advances in computing provide new opportunities for understanding biology, and this will continue to be true for the foreseeable future. Advances in biological understanding may yet have enormous value for changing computing paradigms (e.g., as may be the case if neural information processing is understood more fully)—but these advances are themselves contingent on work done over a considerably longer time scale.

ILLUSTRATIVE PROBLEM DOMAINS AT THE BIOCOMP INTERFACE

Both life scientists and computer scientists will draw inspiration and derive utility from other fields—including each other's—as they see fit. Nevertheless, one way of making progress is to address problems that emerge naturally at the BioComp interface. Problem-focused research carries the major advantage that problems offered by nature do not respect disciplinary boundaries; hence, in making progress against challenging problems, practitioners of different disciplines must learn to work on problems that are shared.

The BioComp interface drives many problem domains in which the expenditure of serious intellectual effort can reasonably be expected to generate significant new knowledge in biology and/or computing. Compared to many of grand challenges in computational biology outlined over the past two decades, making significant progress in these problem domains will call for a longer time scale, greater resources, and more extensive basic progress in computing and in biology.

Biological insight could take different forms—the ability to make new predictions, the understanding of some biological mechanism, the construction of a synthetic biological mechanism. The same is true for computing—insight might take the form of a new biologically inspired approach to some computing problem, different hardware, or novel architecture.

This report discusses a number of interesting problem domains at the BioComp interface, but given the breadth of the cognizant scientific arenas, no attempt is made to be exhaustive. Rather, topics have been selected to span a space of possible problem domains, and no inferences should be made concerning the omission of any problem from this list. The problem domains discussed in this report include high-fidelity cellular modeling and simulation, the development of a synthetic cell, neural information processing and neural prosthetics, evolutionary biology, computational ecology, models that facilitate individualized medicine, a digital human on which a surgeon can operate virtually, computational theories of self-assembly and self-modification, and a theory of biological information and complexity.

THE ROLE OF ORGANIZATION AND INFRASTRUCTURE IN CREATING OPPORTUNITIES AT THE INTERFACE

The committee believes that over time, computing will assume an increasing role in the working lives of nearly all biologists. But given the societal benefits that accompany a fuller and more systematic understanding of biological phenomena, it is better if the computing-enabled 21st century biology arrives sooner rather than later.

This point suggests that cultural and organizational issues have at least as much to do with the nature and scope of the biological embrace of computing as do intellectual ones. The report discusses barriers to cooperation arising from differences in organizational culture and differences in intellectual style.

Consider organizational cultures. In many universities, for example, it is difficult for scholars working at the interface between two fields to gain recognition (e.g., tenure, promotion) from either—a fact that tends to drive such individuals toward one discipline or another. The short-term goals in industrial settings also inhibit partnerships along the interface because of the longer time frame for payoff. Nonetheless, the committee believes that a synergistic cooperation between practitioners in each field, in both basic and applied settings, will have enormous payoffs despite the real differences in intellectual style.

Coordination costs are another issue, because they increase with interdisciplinary work. Computer scientists and biologists are likely to belong to different departments or universities, and when they try to work together, the lack of physical proximity makes it harder for collaborators to meet, to coordinate student training, and to share physical resources. In addition, bigger projects increase coordination costs, and interdisciplinary projects are often larger than unidisciplinary projects. Such costs are reflected in delays in project schedules, poor monitoring of progress, and an uneven distribution of information and awareness of what others in the project are doing. They also reduce people's willingness to tolerate logistical problems that might be more tolerable in their home contexts, increase the difficulty of developing mutual regard and common ground, and can lead to more misunderstandings.

Differences of intellectual style occur because the individuals involved are first and foremost intellectuals. For example, for the computer scientist, the notions of modeling systems and using abstractions are central to his or her work. Using these abstractions and models, computer scientists are able to build some of the most complex artifacts known. But many—perhaps most—biologists today have a deep skepticism about theory and models, at least as represented by mathematics-based theory and computational models. And many computer scientists, mathematicians, and other theoretically inclined researchers fail to recognize the complexity inherent in biological systems. As a result, there is often an intellectual tension between simplification in service of understanding and capturing details in service of fidelity—and such a tension has both positive and negative consequences.

Cooperation will require that practitioners in each field learn enough about the other to engage in substantive conversations about hard biological problems. To take one of the most obvious examples, the different fields place different emphases on the role of empirical data vis-à-vis theory. Accurate data from biological organisms impose "hard" constraints on the biologist in much the same way that results from theoretical computer science impose hard constraints on the computer scientist. A second example is that whereas computer scientists are trained to develop general solutions that give guarantees about events in terms of their worst-case performance, biologists are interested in specific solutions that relate to very particular (though voluminous) datasets.

Finally, institutional difficulties often arise in academic settings for work that is not traditional or not easily identified with existing departments. These differences derive from the structure and culture of departments and disciplines, and they lead to scientists in different disciplines having different intellectual and professional goals and experiencing different conditions for their career success. Collaborators from different disciplines must find and maintain common ground, such as agreeing on goals for a joint project, but must also respect one another's separate priorities, such as having to publish in primary journals, present at particular conferences, or obtain tenure in their respective departments according to departmental criteria. Such cross-pressures and expectations from home departments and disciplinary colleagues remain even if the participants in a collaboration develop similar goals for a project.

FINDINGS AND RECOMMENDATIONS

At the outset, the committee had hoped to identify a deep symmetry between computing and biology. That is, it is clear that the impact of computing on biology is increasingly profound, and the symmetrical notion would be that biology would have a comparable effect on computing. However, this proved not to be the case. The impact of computing on biology will be deep and profound, and indeed will span virtually all areas of life sciences research, and in this direction a focus on interesting problem domains (some of which are illustrated above) is a reasonable way to proceed. By contrast, research that explores the impact of biology on computing falls much more into the "high-risk, high-payoff" category. That is, the ultimate value of biology for changing computing paradigms in deep and fundamental ways is as yet unproven. Nevertheless, various biological attributes—robustness, adaptation, damage recovery, and so on—are so desirable from a computing point of view that any intellectual inquiry is valuable if it can contribute to human-engineered artifacts with these attributes.

It is also clear that a number of other areas of inquiry are associated with the BioComp interface; in addition to biology and computing, the interface also draws from chemistry, materials science, bioengineering, and biochemistry. Three of the most important efforts, which can be loosely characterized as different flavors of biotechnology, are (1) analytical biotechnology (which involves the application of biotechnological tools for the creation of chemical measurement systems); (2) materials biotechnology (which entails the use of biotechnological methods for the fabrication of novel materials with unique optical, electronic, rheological, and selective transport properties); and (3) computational biotechnology (which focuses on the potential replacement of silicon devices with nanoscale biomolecular-based computational systems).

The committee underscores the importance of building human capital and, within that enterprise, the special significance of educational innovation at the BioComp interface. The committee endorses the call from other reports that recommend greater training in quantitative sciences (e.g., mathematics, computer sciences) for biologists, but it also believes that students of the new biology would benefit greatly from some study of engineering. Just as engineers must construct physical systems to operate in the real world, so also must nature operate under these same constraints—physical laws—to "design" successful organisms. Despite this fundamental similarity, biology students rarely learn the important analysis, modeling, and design skills common in engineering curricula. The committee believes that the particular area of engineering (electrical, mechanical, computer, etc.) is probably much less relevant than exposure to essential principles of engineering design: the notion of trade-offs in managing competing objectives, control systems theory, feedback, redundancy, signal processing, interface design, abstraction, and the like.

Of course, more than education will have to change. Fifty years ago, academic biology had to choose between altering the then-dominant styles of research to embrace molecular biology or risking obsolescence. The committee believes that a new dawn is visible—and just as molecular biology has become simply part of the biological sciences as a whole, so also will computational biology ultimately become simply a part of the biological sciences. In the interim, however, considerable effort will be required to build and sustain the infrastructure and to train a generation of biologists and computer scientists who can choose the right collaborators to thrive at the BioComp interface.

The committee believes that 21st century biology will be based on a synergistic mix of reductionist and systems biologies. For systems biology researchers, the committee emphasizes that empirical and experimental hypothesis-testing research will continue to be central in providing experimental verification of putative discoveries—and indeed, relevant as much to studies of how components interact as to studies of components themselves. Thus, disparaging rhetoric about the inadequacies and failures of reductionist biology and overheated zeal in promoting systems biology should be avoided. For researchers more oriented toward experimental or empirical work, the committee emphasizes that systems biology will be central in formulating novel, interesting, and in some cases counterintuitive hypotheses to test. The point suggests that agencies that have traditionally supported hypothesis-testing research would do well to cast a wide "discovery" net that supports the development of alternative hypotheses as well as research that supports traditional hypothesis testing.

Twenty-first century biology will require leadership from both biology and computing that links together first-class research efforts in their respective domains. These efforts will necessarily cross traditional institutional boundaries. For example, research efforts in scientific computing will have to exist in both clinical and biological environments if they are to couple effectively to problem domains in the life sciences. Establishment of a pervasive national infrastructure for life sciences research (including the construction of interdisciplinary teams) and development of the requisite IT-enabled tools for the larger community will require both sustained funding and rigorous oversight. Likewise, the departmental imperatives that characterize much of academe will have to be modified if work at the BioComp interface is to flourish.

In general, the committee believes that the most important change in funding policy for the supporters of this area would be to broaden the kinds of work for which they offer support to include the development of technology for data acquisition and analysis and exploratory research that results in the generation of interesting hypotheses to be tested. That said, there is a direct relationship between the speed with which research frontiers advance and the levels of funding allocated to them. Although it understands the realities of a budget-constrained environment, the committee would gladly endorse an increased flow of funding to the furtherance of a truly integrated 21st century biology.

As for the support of biologically inspired computing, the committee believes that its high-risk, high-payoff nature means that supporting agencies should take a broad view of what "biological inspiration" means and should support the field on a level-of-effort basis, recognizing the long-term nature of such work and taking into account the number of researchers doing and likely to do good work in this area and the potential availability of other avenues to improved computing.

From the committee's perspective, the high-level goals articulated by the agencies and programs that support work related to biology's potential contribution to computing seem generally sensible. This is not to say that every proposal supported under the auspices of these agencies' programs would necessarily have garnered the support of the committee—but that would be true of any research portfolio associated with any program.

One important consequence of supporting high-risk research is that it is unlikely to be successful in the short term. Research—particularly of the high-risk variety—is often more "messy" and takes longer to succeed than managers would like. Managers understandably wish to terminate unproductive lines of inquiry, especially when budgets are constrained. But short-term success cannot be the only metric of the value of research, because when it is, funding managers invite hyperbole and exaggeration on the part of proposal submitters, and unrealistic expectations begin to characterize the field. Those believing the hyperbole (and those contributing to it as well) thus overstate the importance of the research and its centrality to the broader goal of improving computing. When unrealistic expectations are not met (and they will not be met, almost by definition), disillusionment sets in, and the field becomes disfavored from both a funding and an intellectual standpoint.

From this perspective, it is easy to see why support for certain fields rises rapidly and then drops precipitously. Wild budget fluctuations and an unpredictable funding environment that changes goals rapidly can damage the long-term prospects of a field to produce useful and substantive knowledge. Funding levels do matter, but programs that provide steady funding in the context of broadly stated but consistent intellectual goals are more likely to yield useful results than those that do not.

Thus, the committee believes that in the area of biologically inspired computing, funding agencies should have realistic expectations, and these expectations should be relatively modest in the near term. Intellectually, their programs should continue to take a broad view of what "biological inspiration" means. Funding levels in these areas ought to be established on a level-of-effort basis (i.e., what the agency believes is a reasonable level of effort to be expended in this area), by taking into account the number of researchers doing and likely to do good work in an area and the potential availability of other avenues to improved computing. In addition, programmatic continuity for biologically inspired computing should be the rule, with playing rules and priorities remaining more or less constant in the absence of profound scientific discovery or technology advances in the area.

CLOSING THOUGHTS

The impact of computing on biology can fairly be considered a paradigm change as biology enters the 21st century. Twenty-five years ago, biology saw the integration of multiple disciplines from the physical and biological sciences and the application of new approaches to understand the mechanisms by which simple bacteria and viruses function. The impact of the early efforts was so significant that a new discipline, molecular biology, emerged, and many biologists, including those working at the level of tissues or systems and whole organisms, came to adopt the approaches and even often the techniques. Molecular biology has had such success that it is no longer a discipline but simply part of life sciences research itself.

Today, the revolution lies in the application of a new set of interdisciplinary tools: computational approaches will provide the underpinning for the integration of broad disciplines in developing a quantitative systems approach, an integrative or synthetic approach to understanding the interplay of biological complexes as biological research moves up in scale. Bioinformatics provides the glue for systems biology, and computational biology provides new insights into key experimental approaches and how to tackle the challenges of nature. In short, computing and information technology applied to biological problems is likely to play a role for 21st century biology that is in many ways analogous to the role that molecular biology has played across all fields of biological research for the last quarter-century—and computing and information technology will become embedded within biological research itself.